Published in final edited form as:
Stat Med. 2012 December 20; 31(29): 3885–3895. doi:10.1002/sim.5477.

A Trivariate Continual Reassessment Method for Phase I/II Trials of Toxicity, Efficacy, and Surrogate Efficacy

Wei Zhong, **Joseph S. Koopmeiners**, and **Bradley P. Carlin**[1]

[1]Wei Zhong is Graduate Assistant, Joseph S. Koopmeiners is Assistant Professor, and Bradley P. Carlin is Professor and Head, Division of Biostatistics, University of Minnesota

Abstract

Recently, many Bayesian methods have been developed for dose-finding when simultaneously modeling both toxicity and efficacy outcomes in a blended phase I/II fashion. A further challenge arises when all the true efficacy data cannot be obtained quickly after the treatment, so that surrogate markers are instead used (e.g, in cancer trials). We propose a framework to jointly model the probabilities of toxicity, efficacy and surrogate efficacy given a particular dose. Our trivariate binary model is specified as a composition of two bivariate binary submodels. In particular, we extend the bCRM approach [1], as well as utilize the Gumbel copula of Thall and Cook [2]. The resulting trivariate algorithm utilizes all the available data at any given time point, and can flexibly stop the trial early for either toxicity or efficacy. Our simulation studies demonstrate our proposed method can successfully improve dosage targeting efficiency and guard against excess toxicity over a variety of true model settings and degrees of surrogacy.

Keywords

Bayesian adaptive methods; Continual reassessment method (CRM); Maximum tolerated dose (MTD); Phase I/II clinical trial; Surrogate efficacy; Toxicity

1 Introduction

In traditional phase I clinical trials, we seek the maximum tolerated dose (MTD) of an investigational agent, which represents the highest dose with toxicity probability less than a physician-specified acceptable maximum. Based on the estimated MTD from a phase I trial, a phase II trial may be conducted to test the agent's efficacy and possibly refine the optimal dosage for futher studies. There are two general classes of phase I clinical trial designs based on dose assignment: rule-based designs, and model-based designs [3]. Rule-based designs include pharmacologically two-stage designs [4], and the traditional 3+3 design [5] and its variations. Among model-based designs, the continual reassessment method, or CRM [6], is a Bayesian design that has been repeatedly shown to have better operating characteristics than rule-based designs. The CRM links the true toxicity probabilities and dose levels through a simple one-parameter model, and updates the posterior estimates of the MTD arising from this parameter continuously as patients are enrolled.

A limitation of the preceding designs is that efficacy is ignored and dose-finding is based only on toxicity. This is problematic in settings where a dose exists that is less than the MTD but further escalation would not result in increased efficacy. Therefore, dose-finding which incorporates both toxicity and efficacy in a blended phase I/II fashion might be a better strategy. With this goal in mind, many statistical methods have been developed for simultaneously modeling both toxicity *and* efficacy outcomes.

1.1 The bCRM method and its extension

We now briefly mention a few recent extensions to the CRM, especially those designed to handle multiple outcomes. Braun [1] extended to a bivariate CRM (bCRM) design by constructing a conditional probability model for both efficacy and toxicity. Thall and Cook [2] jointly modeled efficacy and toxicity using a bivariate Gumbel copula [7] and introduced an efficacy-toxicity trade-off contour in two-dimensional space to guide dosage selection. Bekele and Shen [8] established a probit model with latent variables to jointly investigate a binary and a continuous outcome. Zhang et al. [9] extended the CRM to a "TriCRM" which is actually univariate but redefines the bivariate binary outcome to a trivariate one: no efficacy or toxicity, efficacy without toxicity, and toxicity with or without efficacy. Nonparametric dose-finding methods have also been proposed to avoid the rigid functional form between the dose and true probability of toxicity or efficacy; see for example Gasparini and Eisele [10] and Yin et al. [11].

Phase I designs that consider efficacy and toxicity rely on the timely availability of the efficacy and toxicity outcomes. In practice, it is not uncommon for toxicity to be available in a short timeframe, while a relatively long wait is required to observe efficacy. A possible solution to this problem is the use of surrogate markers as end points for efficacy. A surrogate marker is defined as a "laboratory measurement or physical sign that is used in therapeutic trials as a substitute for a clinically meaningful endpoint, that is a direct measure of how a patient feels, functions, or survives and is expected to predict the effect of the therapy" [12]. For instance, in cancer trials efficacy may be defined as a survival outcome after some fairly long period of time, say 5 years. In such cases, surrogate markers, such as a seriological measure a physician might check soon after treatment administration, are commonly used for guiding decisions about whether the intervention is promising enough to justify a large definitive trial with clinically meaningful outcomes [13].

A fairly common situation in oncology is when toxicity and surrogate efficacy may be obtained quickly, whereas only part or none of the final efficacy data might be available. One possible solution is to simply replace the missing efficacy data with the surrogate efficacy data in our analysis. However, this could be misleading in the case of a weak surrogate; the mixture of the efficacy and surrogate efficacy data might lead to incorrect conclusions since the primary and surrogate endpoints may react quite differently to the drug. In such settings, more sophisticated biostatistical methods are needed to properly utilize all available information.

1.2 Motivating Example

We motivate our design by a Phase I dose-escalation study to evaluate a novel NK cell treatment for patients with non-Hodgkin lymphoma. The goal of this study is to evaluate the safety of the treatment, and to identify an optimal dose for further evaluation in Phase II. Unlike standard chemotherapeutic agents, here there may be a dose such that further escalation would increase the probability of toxicity without a corresponding increase in the probability of efficacy. In this sense, the optimal dose for further investigation in Phase II is likely to be less than the MTD, and we must consider the tradeoff between efficacy and toxicity during dose escalation.

As is typical in phase I, efficacy and toxicity are evaluated as dichotomous outcomes. In our case, toxicity is defined as any grade 3 or higher toxicity during the first 6 weeks, and efficacy is defined as tumor response at week 15. The timing of these outcomes poses an obvious problem. The 9-week delay between evaluation of the toxicity and efficacy endpoints would delay enrollment of new cohorts, and increase the overall length of our study to the point where its design is no longer practical.

Fortunately, in addition to the efficacy and toxicity outcomes, absolute lymphocyte count will be measured at week 2. Absolute lymphocyte count responds quickly to NK cell treatments and is often used as a surrogate for treatment response. While such early-phase surrogates are sometimes unreliable, an absolute lymphocyte count greater than 1000 cells/μl is thought to predict tumor response at 15 weeks. One approach to overcoming the long delay between the evaluation of toxicity and the evaluation of efficacy is thus to consider the surrogate endpoint in place of the efficacy endpoint, but this could lead to incorrect conclusions about the efficacy of our drug if the surrogate endpoint does not predict true efficacy as well as anticipated. Ideally, we would prefer a dose-escalation study that makes use of the surrogate endpoint but also incorporates efficacy information as it becomes available.

In an adaptive fashion, we design a trial that enrolls a new cohort with sample size $c = 3$ every 6 weeks. This implies that at the enrollment time of the mth ($m \geq 3$) cohort, the efficacy data of the $(m-2)th$ and $(m-1)th$ cohorts are unavailable. The maximum number of cohorts, L, is set to 11. Figure 1 shows the patient enrollment plan as well as how to visualize the data available at a particular enrollment time (here, week 18).

In this paper, we propose a framework to jointly model the probabilities of toxicity, efficacy and surrogate efficacy given a specific dose. Our trivariate binary model is specified as a composition of two bivariate binary submodels. In particular, we extend the bCRM approach, as well as utilize the Gumbel copula of Thall and Cook [2]. The full Bayesian design consists of three elements: a trivariate binary model, a set of sensible prior distributions, and a dose-finding algorithm. Given these elements, we can repeatedly generate artificial data from our design, and thus simulate its (Bayesian or frequentist) operating characteristics, notably the empirical probabilities of correct dose selection, and the proportions of trial participants assigned to each dose.

The rest of the article is organized as follows. In Section 2 we present our general framework for the trivariate probability model, and our two preferred parametric model specifications. Bayesian prior selection and a dose-finding algorithm are addressed as well. Section 3 presents our simulation results under different scenarios, and provides guidance on specifying a future design. Finally, in Section 4 we discuss the strengths and limitations of our proposed method, and discuss areas for further investigation.

2 Methods

2.1 Trivariate Probability Model

Let $Y_{ij} = (T_{ij}, S_{ij}, E_{ij})$ be the binary indicators of toxicity (T), surrogate efficacy (S), and efficacy (E) for subject i who receives a drug treatment at dose X_j. The joint trivariate distribution can be decomposed into three parts as

$$Pr(T_{ij}=t, E_{ij}=e, S_{ij}=s)=Pr(T_{ij}=t)Pr(E_{ij}=e|T_{ij}=t)Pr(S_{ij}=s|T_{ij}=t, E_{ij}=e), \quad (1)$$

where t, e and $s \in \{0, 1\}$. If we assume conditional independence of S and T given E, the above joint distribution can be simplified to

$$Pr(T_{ij}=t, E_{ij}=e, S_{ij}=s)=Pr(T_{ij}=t)Pr(E_{ij}=e|T_{ij}=t)Pr(S_{ij}=s|E_{ij}=e). \quad (2)$$

Although this conditional independence assumption may initially seem a bit strong, note that we do not *assume* marginal independence of S and T; only that their correlation is accounted for by E. Moreover, our model is consistent with the following latent process that might plausibly generate the trivariate binary outcomes:

Toxicity ⇒ Efficacy ⇒ Surrogate efficacy

Biologically, we can think of the data arising from two correlated latent processes, one corresponding to toxicity and the other to efficacy. Here, the efficacy and surrogate efficacy endpoints are both generated by the latent process for efficacy. Our conceptual model assumes that the correlation between the two latent processes is fully captured by the conditional probability of efficacy given toxicity, in which case it is reasonable to assume that surrogate efficacy is independent of toxicity given efficacy. In this way, the trivariate joint distribution can be represented as a product of one marginal distribution for T and two conditional distributions for E and S respectively. Therefore we can flexibly apply various parametric link functions for the marginal or conditional submodels, as we now describe.

2.2 Parametric Functions of Submodels

To monitor the marginal probabilities of T, E and S (p_{Tj}, p_{Ej} and p_{Sj} respectively) given dose X_j, we apply three simple logistic regression models as follows:

$$log\left(\frac{p_{Tj}}{1-p_{Tj}}\right)=\alpha_T+\beta_T X_j, \quad (3)$$

$$log\left(\frac{p_{Ej}}{1-p_{Ej}}\right)=\alpha_E+\beta_E X_j, \quad (4)$$

$$\text{and } log\left(\frac{p_{Sj}}{1-p_{Sj}}\right)=\alpha_S+\beta_S X_j, \quad (5)$$

where α_T, α_S and α_E are assumed to be negative to account for the small probabilities, while β_T, β_S and β_E are assumed to be positive, since both efficacy and toxicity are assumed certain to be increasing with dose. This parametric model is commonly seen in many phase I designs, which assumes a monotonic relationship between dose levels and outcomes.

It is more complicated once we involve the two conditional probabilities ($Pr(E_{ij}=e \mid T_{ij}=t)$ and $Pr(S_{ij}=s \mid E_{ij}=e)$). For a bivariate binary distribution, we investigate approaches endorsed by Braun [1] and Thall and Cook [2], respectively. Suppose Z_1 and Z_2 are two binary random variables, or z_1, $z_2 \in \{0, 1\}$. Based on the work of Arnold and Strauss [14], Braun [1] suggests the following copula model for the conditional probability of $Z_1 = 1$ given $Z_2 = z_2$, namely *a* Bernoulli with success probability

$$Pr(Z_1=1|Z_2=z_2)=\frac{p_1\phi^{z_2}(1-\phi)^{(1-z_2)}}{p_1\phi^{z_2}(1-\phi)^{(1-z_2)}+(1-p_1)(1-\phi)}. \quad (6)$$

Note this reduces to p_1 when $Z_2 = 0$ and $\frac{p_1\phi}{p_1\phi+(1-p_1)(1-\phi)}$ when $Z_2 = 1$. Here, ϕ captures the association between Z_1 and Z_2, with $\phi=\frac{1}{2}$ implying independence between Z_1 and Z_2, $\phi>\frac{1}{2}$ indicating positive association, and $\phi<\frac{1}{2}$ indicating negative association. Note that (6) equals p_1 for all Z_2 where $\phi=\frac{1}{2}$, clarifying the independence case. A drawback to this specification is that the joint bivariate distribution of Z_1 and Z_2 is not available as a standard family. Still, this model specification can often lead to trial designs with good operating

characteristics. To adapt it to our framework, we need only set Z_1 as E and Z_2 as T to establish $Pr(E|T)$, say with association parameter ϕ_1. Similarly, an association parameter ϕ_2 can be used in a conditional model for $Pr(S|E)$. This extension of Braun's method is denoted by "ExB"(extended Braun) method in this paper.

Another approach to studying bivariate binary outcomes is the Gumbel copula utilized by Thall and Cook [2]. The joint bivariate Gumbel copula distribution is specified as

$$Pr(Z_1=z_1, Z_2=z_2)=p_1^{z_1}(1-p_1)^{(1-z_1)}p_2^{z_2}(1-p_2)^{(1-z_2)}+(-1)^{z_1+z_2}p_1(1-p_1)p_2(1-p_2)\frac{e^{\gamma}-1}{e^{\gamma}+1}, \quad (7)$$

where p_1 and p_2 are the *marginal* probabilities success for Z_1 and Z_2, respectively; note that their interpretations have changed somewhat from p_1 and p_2 in the Braun model. Similar to ϕ, γ captures the association between Z_1 and Z_2, but now where the range of γ covers the whole real line and $\gamma = 0$ corresponds to independence, which is immediately apparent from (7). Positive values of γ indicate positive association, while negative values imply negative association. Then the conditional probability of $Z_1 = 1$ given $Z_2 = z_2$ can be easily calculated as

$$\begin{aligned} Pr(Z_1=1|Z_2=z_2) &= \frac{Pr(Z_1=1,Z_2=z_2)}{Pr(Z_1=1,Z_2=z_2)+Pr(Z_1=0,Z_2=z_2)} \\ &= p_1+(-1)^{z_2+1}p_1(1-p_1)p_1^{1-z_2}(1-p_2)^{z_2}\frac{e^{\gamma}-1}{e^{\gamma}+1}, \end{aligned}$$

which is distinct from (6); in particular we note the dependence on p_2. In the same fashion as our previous model extension, two association parameters γ_1 and γ_2 can be used in the two required conditional probabilities $Pr(E|T)$ and $Pr(S|E)$. We refer to this extension of the Gumbel copula as the "ExG"(extended Gumbel) method in what follows.

2.3 Likelihood and Prior Specification

As with all Bayesian analyses, a full likelihood and a prior distribution for every parameter are required. Following Braun [1], α_T, α_E and α_S in equations (3), (4) and (5) are all set equal to the constant −3, to reflect the relative rarity of response with the lowest doses of the agent. Suppose that n_j patients are treated at dose X_j (j = 1,…,k), among which n_j^{tes} patients have outcomes $T = t$, $E = e$ and $S = s$. Then the assuming *complete* toxicity, efficacy, and surrogate efficacy data, the likelihood for the "ExB" model would be multinomial,

$$L_C(\beta_T,\beta_S,\beta_E,\phi_1,\phi_2|\text{Data}) \propto \prod_{j=1}^{k} \pi_j^{000^{n_j^{000}}} \pi_j^{001^{n_j^{001}}} \pi_j^{010^{n_j^{010}}} \pi_j^{011^{n_j^{011}}} \pi_j^{100^{n_j^{100}}} \pi_j^{101^{n_j^{101}}} \pi_j^{110^{n_j^{110}}} \pi_j^{111^{n_j^{111}}}, \quad (8)$$

where π_j^{tes} represents the joint probability of T, E, and S given dose X_j, and (ϕ_1, ϕ_2) is replaced by (γ_1, γ_2) in the "ExG" model. The joint probabilities π_j^{tes} can be computed as described in the previous two subsections. However, as illustrated in Figure 1, part or all of the efficacy data are missing at the time of dose assignment, so that n_j^{tes} for all the subjects cannot be fully determined. Assuming the efficacy data of subject i treated at dose j is unavailable at a specific time, the marginal likelihood contribution for this subject is

$$Pr(T_{ij}=t, S_{ij}=s)=\left\{Pr(E_{ij}=1|T_{ij}=t)Pr(S_{ij}=s|E_{ij}=1)+Pr(E_{ij}=0|T_{ij}=t)Pr(S_{ij}=s|E_{ij}=0)\right\}Pr(T_{ij}=t)$$

by (2) and the law of total probability. One could use this to construct a missing data likelihood, but in fact the BUGS language tackles the problem simply by imputing any still-

missing efficacy datum from its full conditional distribution (i.e., the efficacy model) before sampling from the (now complete-data) full conditionals derived from (8), an approach that is mathematically equivalent.

Turning to the priors, we assume β_T, β_S and β_E all independently follow exponential distributions with mean 1. In the "ExB" method, ϕ_1 is assumed to follow a *Beta*(2, 2) distribution and ϕ_2 a *Beta*(4, 2) distribution, which encourages prior independence between E and T, but positive dependence a priori between E and S. In the "ExG" method, the priors for γ_1 and γ_2 are set as normal distributions ($N(0, 5)$ and $N(0.69, 5)$, respectively), priors designed to match the two beta priors as closely as possible, but on the γ scale.

2.4 Dose-Finding Algorithm

Suppose at a specific enrollment time point, Y denotes all the available accumulated data. Let $E[p_{Tj}|Y]$ and $E[p_{Ej}|Y]$ be the posterior mean probabilities of toxicity and efficacy. After each look at the data, Braun [1] suggests updating $E[p_{Tj}|Y]$ and $E[p_{Ej}|Y]$, and then selecting the dose by minimizing

$$dist_j = \sqrt{(E[p_{Tj}|Y] - p_T^*)^2 + (E[p_{Ej}|Y] - p_E^*)^2}, \quad (9)$$

the Euclidean distance between the current estimates and some physician-specified target rates of toxicity and efficacy, p_T^* and p_E^*. Here we propose a few modifications of this basic approach. First, to discourage excessive toxicity caused by high doses, we put higher weight on the toxicity component contribution to (9). Second, we permit different penalties for over and under-dosing by incorporating asymmetry into the distance calculation. Specifically, we let

$$dist_{Tj} = \begin{cases} (E[p_{Tj}|Y] - p_T^*), & \text{if } E[p_{Tj}|Y] > p_T^* \\ w_T(E[p_{Tj}|Y] - p_T^*), & \text{otherwise} \end{cases} \quad (10)$$

$$\text{and } dist_{Ej} = \begin{cases} (E[p_{Ej}|Y] - p_E^*), & \text{if } E[p_{Ej}|Y] > p_E^* \\ w_E(E[p_{Ej}|Y] - p_E^*), & \text{otherwise} \end{cases} \quad (11)$$

where $0 < w_T < 1$ and $0 < w_E < 1$. The full distance for a specific dose X_j is then a modified version of (9), namely

$$dist_j = \sqrt{W_{dT}\, dist_{Tj}^2 + W_{dE}\, dist_{Ej}^2}, \quad (12)$$

where $W_{dT} > 0$ and $W_{dE} > 0$ are positive weights that can be adjusted to achieve better operating characteristics for the trial. A dose with a smaller value of $dist_j$ is more desirable since it is closer to the pre-specified probabilities of toxicity and efficacy. Finally, we also employ termination rules to control overtoxicity and to enable an early decision regarding the optimal dosage. First, if for the lowest dose level X_1, the posterior samples of p_{T1} satisfy

$$Pr(p_{T_1} < p_T^* | Y) < \tilde{\pi}_T,$$

where $\tilde{\pi}_T$ is some pre-specified small value (say, 0.2), then we will terminate the trial for over-toxicity. Second, if for some dose X_j, the posterior samples of p_{Tj} and p_{Ej} satisfy

$$Pr(p_{T_j} < p_T^* | Y) > \dot{\pi}_T, \text{ and } Pr(p_{E_j} < p_E^* | Y) > \dot{\pi}_E,$$

where $\dot{\pi}_T$ and $\dot{\pi}_E$ are two pre-specified large probabilities (say, 0.8), then we will stop the trial and define dose X_j as the optimal dose. If there are multiple doses which satisfy both probability statements, then the dose with the smallest $dist_j$ in (12) is picked as the optimal dose.

In summary, our proposed trivariate dose-finding algorithm is as follows:

Trivariate Dose-finding Algorithm

1. Treat the first cohort patients at the lowest dose level.
2. Update the posterior distributions of the probabilities of toxicity and efficacy at all dose levels.
3. Calculate the criteria to check for early trial termination.
4. If not terminated, calculate the distances $dist_j$ for $j = 1,\ldots,5$.
5. Treat the next patient cohort at the dose having the minimum distance (12) under the restrictions of no dosage shift of more than one level of escalation or deescalation.
6. Repeat from Step 2 until the trial is terminated early or maximum sample size is achieved.

Weight choice depends on the practical situation and the utility function of the investigator, a subject we return to below. Regarding $\tilde{\pi}_T$ reducing it may lead to more early stopping due to toxicity. If researchers are confident that the lowest dose level will not exceed the target level, a larger value of $\tilde{\pi}_T$ can be used. Then $\dot{\pi}_T$ and $\dot{\pi}_E$ can be determined by the desired aggressiveness of the trial design. For small sample size designs, we may be more interested in completing the trial than in stopping it due to good performance, whence larger values of $\dot{\pi}_T$ and $\dot{\pi}_E$ (i.e., close to 1) will be more reasonable.

3 Simulation Results

We now present the result of several simulation studies based on the motivating example in Section 1.2. The quality of surrogate marker in our models can be evaluated in two ways: the difference between S and E in marginal posterior probability, or via the association parameters (ϕ_2 or γ_2) between S and E. We define a “good” surrogate as one that has a strong association and a marginal probability close to that for true efficacy, while a “bad” surrogate has a weak association and a dissimilar marginal probability. To model the false positivity of the surrogate in a real situation, we set $Pr(S = 1) = 1.1 * Pr(E = 1)$ for a “good” surrogate, and $Pr(S = 1) = 1.5 * Pr(E = 1)$ for a “bad” surrogate. Note all of our ϕ_1 and γ_1 settings imply modest positive association between E and T, whereas the “bad” surrogate choices for ϕ_2 and γ_2 assume independence of E and S. We set $w_E = w_T = \frac{1}{3}$ as our penalty reduction for undershooting toxicity and overshooting efficacy in (10) and (11), and set $W_{dT} = 2$ but $W_{dE} = 1$ in (12), thus making toxicity twice as important as efficacy in the distance calculation.

For our simulation, we considered both our trivariate joint models (ExB and ExG) and the corresponding bivariate models (Braun and Gumbel) that simply replaced the efficacy data with the surrogate efficacy data, assuming the latter to be without error (as is sometimes

done in practice). Due to the different meanings of ϕ_1 and γ_1 (and ϕ_2 and γ_2), we primarily seek to compare ExB to Braun and ExG to Gumbel, respectively, to evaluate the benefits of our trivariate model. Each of our simulation studies used 1000 simulated trials, each analyzed by generating two MCMC chains in the WinBUGS software (www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml). We ran 1000 MCMC iterations after a 1000-iteration burn-in period for each chain. Standard convergence diagnostics [15] did not reveal significant MCMC convergence issues.

Table 1 presents parameter settings and target probabilities p_T^* and p_E^* for three different scenarios. Scenario 1 and Scenario 2 assume the optimal doses are Doses 4 and 2, respectively, where here a dose is "optimal" if its true p_T and p_E values respectively equal p_T^* and p_E^*, the physician-supplied target values, and thus correspond to a true distance of 0 in equation (12). The third scenario describes a situation where all doses are over-toxic based on the the physician-supplied target values; that is, $p_{T_j} > p_T^*$ for all *j*. In Scenario 3, none of the five doses correspond to a true distance (12) near 0.

Table 2 shows the empirical selection probabilities and percents of patients treated at each dose for the competing methods under "good" surrogacy across three different scenarios, while Table 3 does the same for "bad" surrogacy. In general, our trivariate models perform better than the corresponding bivariate models, especially under "bad" surrogacy. In addition, using a "good" surrogate offers a better chance of identifying and assigning more patients to the optimal dose than using a "bad" one.

Specifically in Scenario 1, ExB and ExG identify the correct dose (Dose 4) slightly more often than the Braun and Gumbel models under "good" surrogacy, respectively. Both trivariate models are also substantially better under "bad" surrogacy than the corresponding bivariate models. Here the standard Braun and Gumbel models select Dose 4 in only 21% and 28% of the simulated studies, compared to 42% and 49% for ExB and ExG, respectively. The increases in the proportions of patients treated at the correct dose for ExB and ExG relative to Braun and Gumbel model are also more dramatic under "bad" surrogacy (26% vs. 18% and 27% vs. 20%, respectively in Table 3). No early terminations due to over-toxicity were detected in this scenario.

In Scenario 2, comparing ExB and ExG to their corresponding bivariate models (Braun and Gumbel), our proposed trivariate models again have better operating characteristics under the "good" and "bad" surrogacy scenarios. Under "bad" surrogacy, the increases in the probability of selecting the correct dose are again more dramatic, going from 62% with the Braun model to 76% with ExB, and from 64% with the Gumbel model to 78% with ExG. The percentages of patients assigned to the optimal dose is also improved from Braun (48%) and Gumbel (47%) to ExB (59%) and ExG (60%), respectively, under "bad" surrogacy. The stopping probabilities due to over-toxicity are no longer identically 0 in this setting, but very close (typically around 1%).

In Scenario 3, the probability of toxicity for all doses exceeds the physician-specified target (p_E^*=0.4) and the correct decision is to terminate early. The results suggest that our trivariate models can do as well as bivariate models in stopping the trial early due to over-toxicity. In our simulation, all four models successfully stopped the trial early at least 79% of the time due to over-toxicity. In the "good" surrogacy scenario, the average numbers of patients treated under the ExB, Braun, ExG and Gumbel models are 15.1, 15.1, 14.7 and 16.1, respectively, far below the number of the patients in the initial enrollment plan (3*11=33). Note that the average number of patients treated at each dose can also be calculated. For example, the "good" surrogacy results reveal there were $15.1 \times 0.14 \approx 2.1$ patients actually

treated at Dose level 2 on average by the ExB model. Under "bad" surrogacy, the total average numbers of patients treated under the four methods changes only slightly to 15.1, 14.8, 14.7 and 16.2, respectively.

We also considered varying the weights w_T, w_E, W_{dT} and W_{dE}. For example, in Scenario 1, suppose we set $W_{dT} = W_{dE} = 1$, or simply swap the previous values (setting $W_{dT} = 1$ and $W_{dE} = 2$, thus emphasizing efficacy). Then targeting efficiency still improves under our trivariate models (results not shown), but we also observe slightly higher probabilities of overdose (Dose 5) than in our previous version (increasing from 0.16 to 0.28 in the "swapped" version). In practice, we suggest calibrating the weights to the desired level of overdose control. For example, an investigator most interested in controlling toxicity for the trial's patients would take $W_{dT} >> W_{dE}$, and use simulation as we have done to verify acceptable operating characteristics.

Discussion

Our proposed method can successfully improve phase I/II dosage targeting efficiency by jointly modeling toxicity, efficacy and surrogate efficacy. Firstly, whether under "good" or "bad" surrogacy scenarios, the targeting performance is improved by adding some efficacy data, as opposed to using only the surrogate data. The quality of surrogate markers is an important factor in finding an optimal dosage. In the above simulation studies, we assume a higher marginal probability for the surrogate marker, which reflects reality in the use of surrogate markers. When we use only the surrogate efficacy data, as we expect, this makes the final dose more variable, hence a poorer estimate of the optimal dose. Especially under "bad" surrogacy, we modeled a large probability of false positive efficacy, leading to a downward effect in the dose selected. However, with some efficacy data, our joint models can eliminate part of the mean squared error and improve targeting accuracy. We experimented with making the bad surrogate "more bad" by using a 2× multiplier (instead of merely 1.5×) in the true efficacy probabilities. As we expected, we did observe an even bigger detriment to Braun and Gumbel performance, with correct selection probabilities dropping to just 0.08 and 0.05. Under ExB and ExG, these probabilities recover to 0.45 and 0.50.

Secondly, we want to point out that direct comparison of the performance of the ExB and ExG methods might not be sensible, since the interpretation of association parameters ϕ or γ is different in each model and it is thus unclear how fair comparisons can be made. Our results indicates ExG performs quite similar to ExB except that under bad surrogacy in Scenario 1, ExG beats ExG (49% vs 42%). Since ExG and Gumbel models explicitly specify a joint bivariate distribution between outcomes, we would recommend the use of ExG rather ExB.

Thirdly, the purpose of our trivariate model is actually to borrow some strength from the surrogate efficacy data so that we can learn more on the missing efficacy data. Our dose finding methods perform slightly better in a trial with a good surrogate than with a bad surrogate, which is consistent with intuition as well as efforts to find a good surrogate in clinical research. It is reasonable that this improvement is not very strong because the missing efficacy data is not so much. In addition, penalty weights w_T, w_E, W_{dT} and W_{dT} in our models can be flexibly adjusted to obtain an optimal dose under different conditions; we suggest putting more weight on w_T and/or W_{dT} to control over-toxicity.

An alternate approach to overcoming the delay between the measurement of toxicity and the measurement of efficacy is to extend the TITE-CRM (Cheung and Chappell [16]) to accommodate multiple outcomes. A disadvantage of this approach is that it would ignore the

information present in the surrogate outcome. Our trivariate model provides a slight improvement in our ability to correctly identify the optimal dose compared to the bivariate model even when the surrogate is a weak predictor of efficacy. This suggests that including both the surrogate and efficacy endpoints provides additional benefit beyond allowing us to complete the trial in a more reasonable time-frame.

Our setting is just an idealization of actual practice, and could be modified. For example, the use of the rigid logistic function of form with intercept fixed at −3 could be replaced with other parametric or nonparametric forms for the toxicity, efficacy, or surrogate efficacy probabilities [17, 18, 19]. Adding an upper bound $\theta < 1$ on the probability of efficacy (or surrogate efficacy) may also be sensible. We experimented with the addition of a "plateau" parameter that provides an upper bound on efficacy, but found this parameter hard to estimate (since relevant information about it is confined to the rarely-visited highest doses) and led to only very small gains in the performance measures reported in our tables. A quadratic model is also a possibility, but again with our small initial datasets, this approach is not sensible without overly informative priors.

Another extension of our method could be to jointly model several surrogate markers for toxicity and efficacy. All the surrogate markers could be assumed to operate "in parallel," meaning that all their inter-connections are captured conditional on efficacy. As mentioned in Section 2.1, the assumption of conditional independence of toxicity and surrogate efficacy might be too strong. Exploration of the relationship between T, E and S (or multiple Ss) may shed light on this issue. We also tried to detect the association between toxicity, efficacy and surrogate efficacy by posterior samples of related association parameters. Unfortunately, this estimation was not satisfactory, with large posterior variance, perhaps because we are trying to estimate a fairly data-insensitive parameter with a small sample of binary outcomes that themselves contain little information about the association parameter. Finally, our definition of a "good" or "bad" surrogate is a little arbitrary. A better quantification of the quality of surrogacy and its effect on our dose-finding are other subjects for future investigation.

Acknowledgments

The work of the first and last authors was supported in part by NCI grant R01-CA095955.

References

1. Braun T. The bivariate continual reassessment method: extending the CRM to phase I trials of two competing outcomes. Controlled Clinical Trials. 2002; 23(3):240–256. [PubMed: 12057877]
2. Thall P, Cook J. Dose-finding based on efficacy–toxicity trade-offs. Biometrics. 2004; 60(3):684–693. [PubMed: 15339291]
3. Berry, S.; Carlin, B.; Lee, J.; Muller, P. Bayesian Adaptive Methods for Clinical Trials. CRC Press; 2010.
4. Collins J, Grieshaber C, Chabner B. Pharmacologically guided phase I clinical trials based upon preclinical drug development. Journal of the National Cancer Institute. 1990; 82(16):1321. [PubMed: 2143234]
5. Storer B. Design and analysis of phase I clinical trials. Biometrics. 1989; 45(3):925–937. [PubMed: 2790129]
6. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase 1 clinical trials in cancer. Biometrics. 1990; 46(1):33–48. [PubMed: 2350571]
7. Murtaugh P, Fisher L. Bivariate binary models of efficacy and toxicity in dose-ranging trials. Communications in Statistics-Theory and Methods. 1990; 19(6):2003–2020.
8. Bekele N, Shen Y. A Bayesian approach to jointly modeling toxicity and biomarker expression in a phase I/II dose-finding trial. Biometrics. 2005; 61(2):343–354. [PubMed: 16011680]

9. Zhang W, Sargent D, Mandrekar S. An adaptive dose-finding design incorporating both toxicity and efficacy. Statistics in Medicine. 2006; 25(14):2365–2383. [PubMed: 16220478]
10. Gasparini M, Eisele J. A curve-free method for phase I clinical trials. Biometrics. 2000; 56(2):609–615. [PubMed: 10877324]
11. Yin G, Li Y, Ji Y. Bayesian dose-finding in phase I/II clinical trials using toxicity and efficacy odds ratios. Biometrics. 2006; 62(3):777–787. [PubMed: 16984320]
12. Temple R. Are surrogate markers adequate to assess cardiovascular disease drugs? JAMA. 1999; 282(8):790–795. [PubMed: 10463719]
13. Fleming T, DeMets D. Surrogate end points in clinical trials: are we being misled? Annals of Internal Medicine. 1996; 125(7):605. [PubMed: 8815760]
14. Arnold B, Strauss D. Bivariate distributions with conditionals in prescribed exponential families. Journal of the Royal Statistical Society. Series B (Methodological). 1991; 53(2):365–375.
15. Cowles M, Carlin B. Markov chain monte carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association. 1996; 91(434):883–904.
16. Cheung Y, Chappell R. Sequential designs for phase I clinical trials with late-onset toxicities. Biometrics. 2000; 56(4):1177–1182. [PubMed: 11129476]
17. Li Y, Bekele B, Ji Y, Cook J. Dose-schedule finding in phase I/II clinical trials using a bayesian isotonic transformation. Statistics in Medicine. 2008; 27(24):4895–4913. [PubMed: 18563789]
18. Yin G, Yuan Y. A latent contingency table approach to dose finding for combinations of two agents. Biometrics. 2009; 65(3):866–875. [PubMed: 18759848]
19. Mandrekar S, Qin R, Sargent D. Model-based phase I designs incorporating toxicity and efficacy for single and dual agent drug combinations: Methods and challenges. Statistics in Medicine. 2010; 29(10):1077–1083. [PubMed: 20419760]

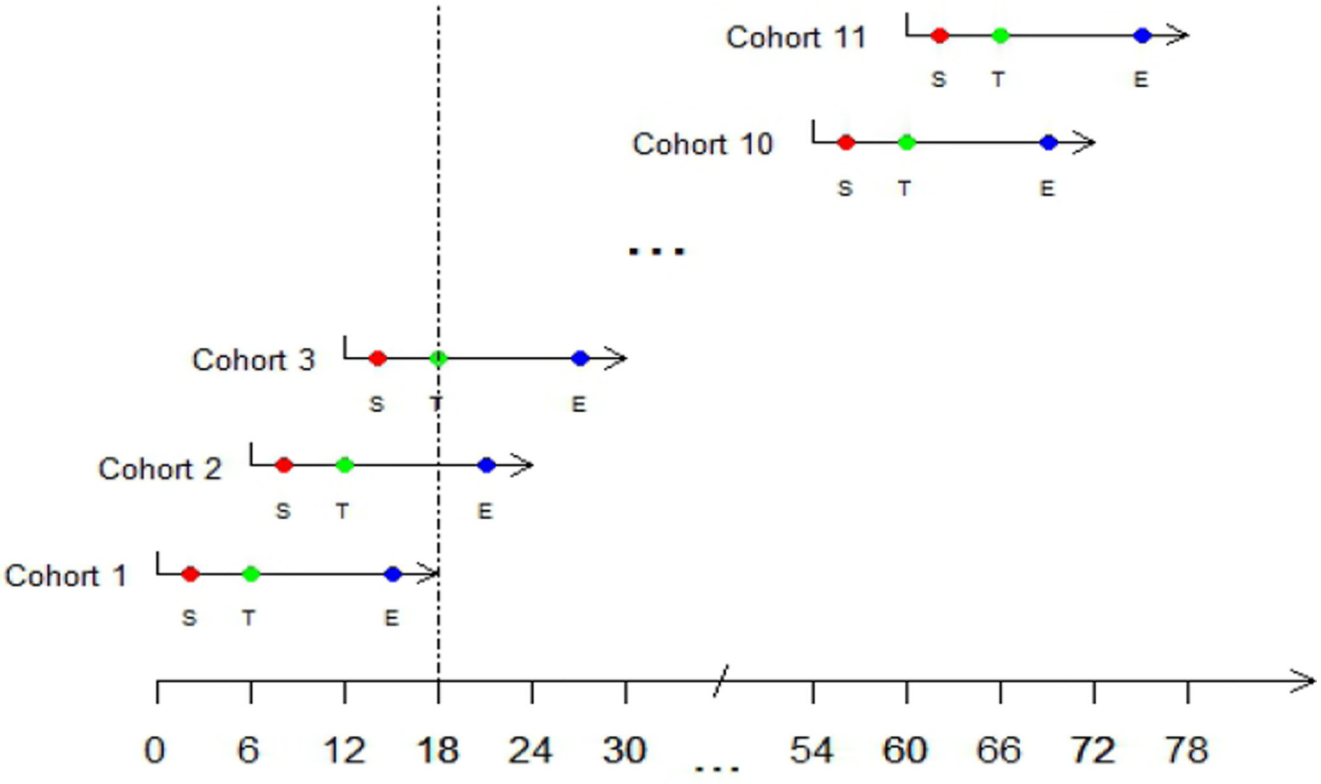

Figure 1.
Patient enrollment schematic for the motivating trial, and the toxicity, efficacy, and surrogate efficacy data available at week 18 (and when deciding the dose assignment for Cohort 4). At this time point, toxicity and surrogate efficacy data are available for all of the first 3 patient cohorts, while efficacy data is available only for Cohort 1.

Table 1

Simulation parameter settings in the three different scenarios. Values for optimal doses are shown in **boldface**.

Dose	1	2	3	4	5
Scenario 1: Dose 4 optimal; target probabilities $p_T^* = 0.27,\ p_E^* = 0.35$					
True Pr(Tox=1)	0.08	0.12	0.18	**0.27**	0.38
True Pr(Eff=1)	0.08	0.14	0.23	**0.35**	0.50
True Pr(Sur=1) good	0.088	0.154	0.253	**0.385**	0.55
True Pr(Sur=1) bad	0.12	0.21	0.345	**0.525**	0.75
True distance	0.28	0.22	0.13	**0**	0.16
Scenario 2: Dose 2 optimal; target probabilities $p_T^* = 0.27,\ p_E^* = 0.35$					
True Pr(Tox=1)	0.12	**0.27**	0.50	0.73	0.88
True Pr(Eff=1)	0.20	**0.35**	0.40	0.48	0.60
True Pr(Sur=1) good	0.22	**0.385**	0.44	0.528	0.66
True Pr(Sur=1) bad	0.30	**0.525**	0.60	0.72	0.9
True distance	0.17	**0**	0.33	0.65	0.87
Scenario 3: All doses over-toxic; target probabilities $p_T^* = 0.4,\ p_E^* = 0.3$					
True Pr(Tox=1)	0.60	0.75	0.80	0.85	0.90
True Pr(Eff=1)	0.08	0.14	0.23	0.35	0.50
True Pr(Sur=1) good	0.088	0.154	0.253	0.385	0.55
True Pr(Sur=1) bad	0.12	0.21	0.345	0.525	0.75
True distance	0.36	0.52	0.57	0.64	0.71
Association parameters:					
ExB (good surrogate)	$\phi_1 = 0.7$	$\phi_2 = 0.9$			
ExB (bad surrogate)	$\phi_1 = 0.7$	$\phi_2 = 0.5$			
ExG (good surrogate)	$\gamma_1 = 1$	$\gamma_2 = 2$			
ExG (bad surrogate)	$\gamma_1 = 1$	$\gamma_2 = 0$			

Table 2

Operating characteristics of the four methods under good surrogacy. Maximal selection probabilities are shown in **boldface**.

			Dose					
Scenario	**Method**	**Operating characteristics**	**1**	**2**	**3**	**4**	**5**	**over-toxic**
Scenario 1: Dose 4 optimal	ExB	selection probability	0.01	0.07	0.26	**0.51**	0.15	0
		proportion of patients treated	0.11	0.20	0.31	0.26	0.12	
	Braun	selection probability	0.03	0.05	0.37	**0.47**	0.08	0
		proportion of patients treated	0.13	0.18	0.33	0.28	0.08	
Scenario 2: Dose 2 optimal	ExB	selection probability	0.13	**0.81**	0.06	0	0	0
		proportion of patients treated	0.24	0.62	0.13	0.01	0	
	Braun	selection probability	0.21	**0.75**	0.03	0	0	0.01
		proportion of patients treated	0.28	0.61	0.10	0.01	0	
Scenario 3: Over-toxic	ExB	selection probability	0.16	0	0	0	0	**0.84**
		proportion of patients treated	0.85	0.14	0.01	0	0	
	Braun	selection probability	0.21	0	0	0	0	**0.79**
		proportion of patients treated	0.86	0.13	0.01	0	0	
Scenario 1: Dose 4 optimal	ExG	selection probability	0	0.04	0.28	**0.52**	0.16	0
		proportion of patients treated	0.11	0.16	0.30	0.31	0.12	
	Gumbel	selection probability	0	0.04	0.29	**0.50**	0.17	0
		proportion of patients treated	0.12	0.16	0.29	0.30	0.13	
Scenario 2: Dose 2 optimal	ExG	selection probability	0.14	**0.80**	0.06	0	0	0
		proportion of patients treated	0.26	0.59	0.14	0.01	0	
	Gumbel	selection probability	0.17	**0.77**	0.05	0	0	0.01
		proportion of patients treated	0.28	0.58	0.13	0.01	0	
Scenario 3: Over-toxic	ExG	selection probability	0.15	0	0	0	0	**0.85**
		proportion of patients treated	0.87	0.13	0	0	0	
	Gumbel	selection probability	0.15	0	0	0	0	**0.85**
		proportion of patients treated	0.87	0.12	0.01	0	0	

Table 3

Operating characteristics of the four methods under bad surrogacy. Maximal selection prob-abilities are shown in **boldface**.

			Dose					
Scenario	**Method**	**Operating characteristics**	**1**	**2**	**3**	**4**	**5**	**over-toxic**
Scenario 1: Dose 4 optimal	ExB	selection probability	0.01	0.06	0.32	**0.42**	0.19	0
		proportion of patients treated	0.12	0.20	0.33	0.26	0.11	
	Braun	selection probability	0.05	0.14	0.59	**0.21**	0.01	0
		proportion of patients treated	0.16	0.22	0.41	0.18	0.03	
Scenario 2: Dose 2 optimal	ExB	selection probability	0.17	**0.75**	0.07	0	0	0.01
		proportion of patients treated	0.26	0.59	0.14	0.01	0	
	Braun	selection probability	0.35	**0.62**	0.02	0	0	0.01
		proportion of patients treated	0.47	0.48	0.05	0	0	
Scenario 3: Over-toxic	ExB	selection probability	0.16	0	0	0	0	**0.84**
		proportion of patients treated	0.85	0.14	0.01	0	0	
	Braun	selection probability	0.20	0	0	0	0	**0.80**
		proportion of patients treated	0.87	0.12	0.01	0	0	
Scenario 1: Dose 4 optimal	ExG	selection probability	0	0.05	0.28	**0.49**	0.18	0
		proportion of patients treated	0.11	0.20	0.27	0.27	0.15	
	Gumbel	selection probability	0	0.09	0.62	**0.28**	0.01	0
		proportion of patients treated	0.12	0.24	0.41	0.20	0.03	
Scenario 2: Dose 2 optimal	ExG	selection probability	0.14	**0.78**	0.07	0	0	0.01
		proportion of patients treated	0.24	0.60	0.15	0.01	0	
	Gumbel	selection probability	0.34	**0.64**	0.01	0	0	0.01
		proportion of patients treated	0.49	0.47	0.04	0	0	
Scenario 3: Over-toxic	ExG	selection probability	0.15	0	0	0	0	**0.85**
		proportion of patients treated	0.89	0.11	0	0	0	
	Gumbel	selection probability	0.16	0	0	0	0	**0.84**
		proportion of patients treated	0.87	0.12	0.01	0	0	